UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Statistical Models II

2122-2-F8203B042-F8203B013M

Learning objectives

The course aims are to provide analytic and inferential advanced statistical procedures also conducted by simulations. The content is presented to develop a critical understanding of the underlying assumptions.

Knowledge and understanding

This teaching allows the student to:

- analyze data with statistical models developed for both categorical and continuous response variables;

- develop simulations independently;

- use the semantics of the R software also through the RMarkdown environment to create reproducible documents containing the code, results, and comments of the analysis;

- rigorously interpret the elaborations' results and provide a detailed and precise description of the same for dissemination purposes.

Ability to apply knowledge and understanding

The course allows the student to:

- develop statistical inference using modern bootstrap techniques;

- estimate, select, and interpret models of mixtures of distributions for heterogeneous populations;

- fit models to latent variables;

- apply theoretical knowledge to the analysis of data collected in various fields, including epidemiology, medicine, biology, genetics, and public health;

- develop code in the R environment.

The student is encouraged to explain the theory and the results by providing written text and oral presentation.

The course provides the main concepts for parametric and non-parametric statistical models that are essential for analyzing the data arising in the working contexts of biostatistics, statistics, demography, and public health. It is compulsory for the next course of student' studies.

Contents

In the first part of the course simulation, methods are introduced to generate pseudo-realizations from random variables. Next, the student is introduced to some resampling methods: bootstrap and jackknife, along with their inferential purposes.

The Expectation-Maximization (EM) algorithm is illustrated for incomplete-data problems through the estimated parameters of a generalized linear model. Then it is illustrated as an optimization method for the estimation model parameters of the finite mixture and latent variable models. The course provides skills in the use of the semantic of the software R.

Detailed program

The first part of the course deals with simulation methods and linear congruential methods to generate pseudo-random numbers. Graphical tools for testing the series are illustrated along with some statistical tests such as Kolmogorov-Smirnov and Chi-Squared tests. Transformations of uniform deviate and simulation of random numbers from specific distributions are considered. Some theoretical features of the exponential, binomial, and Gaussian distributions and convolution of random variables are exposed.

In the second part of the course, the main resampling methods such as the jackknife and bootstrap are introduced. The bootstrap is applied for bias adjustment and the estimation of dispersion. Bootstrap confidence intervals based on the percentile method and the bias-corrected accelerated bootstrap method are explained.

Among the optimization methods, the Expectation-Maximization Algorithm is considered and explained first as a tool to impute missing values through a generalized linear model and then as a tool to maximize the log-likelihood function for incomplete data problems. Finite mixture models are introduced both for continuous and categorical data and a particular focus is given on the mixture of Gaussian distributions and latent variable models for categorical data.

Some time is devoted to explaining the theory by imparting the flavor of the empirical applications on real data collected from different fields arising in epidemiology, pharmacoepidemiology medicine and biology, ecology, and environmental sciences. They are developed within the statistical software R, RStudio with the RMarkdown interface to provide live code develop comments on the results of analyses critically, including through cooperative learning. The main R packages used are: boot, bootstrap MultiLCIRT, and muclust.

Prerequisites

For an **easier** understanding of the course content, it is recommended to know Probability and Statistical Inference notions. The student should also know the basic semantics of the programming language in the R environment.

Teaching methods

The lectures are held in the lab since the theoretical part is placed side by side with the applications carried out with the computer. Many practical examples based on real and simulated data referred to different contexts are proposed to the students to be solved with R through the RMarkdown interface during the lectures. The student is also encouraged to develop cooperative learning to interact with each other and finalize the required steps of the analysis. Exercises are carried out to report in a written form the results by adding critical comments and create reproducible documents.

During the Covid-19 emergency period, classes will be held remotely (videotaped lectures) with periodic meetings in videoconference via Webex platform and/or in-person according to the schedules provided by the university, and that will be announced on the course page.

Assessment methods

The following methods of verifying learning apply to both students attending and non-attending lectures in presence. The examination is in written form with optional oral; there are no intermediate tests. The written exam has a maximum total duration of two hours and takes place in the computer lab. During the test, it is necessary to solve the exercises applied in the light of the theoretical arguments developed during the course and answer some theory questions. The analyses are conducted using the R environment, Rstudio, and RMarkdown. The exercises allow verifying the ability to understand the problem and its resolution by applying advanced statistical models to real or simulated data and the elaboration of reports in which the procedure is described, and the results are illustrated. In addition, the theory questions allow verifying the learning of the theoretical concepts taught during the course.

During the emergency period due to the Covid-19 pandemic depending on the university arrangements the exam will take place in the computer lab or via video conferencing via the Webex platform.

Textbooks and Reading Materials

The teaching material consists mainly of handouts prepared by the teacher. They cover both the theory topics and the applications developed with R or SAS software. All the files are available on the page of the e-learning platform of the university dedicated to the course. In addition, the teacher publishes at the end of each lesson: the slides, the calculation programs, the exercises, the datasets, and the solutions of the exercises. On the same page are also published some previous exam texts.

During the Covid-19 emergency period, video recordings of lectures are also posted on the course page.

The primary reference texts are listed in the bibliography of the handouts (some of these are also available in

ebook); among others the following are noted:

Bartolucci, F., Farcomeni, A., Pennoni, F. (2013). *Latent Markov Models for longitudinal data*, Chapman and Hall/CRC, Boca Raton.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (2007). *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media, New York.

Blitzstein J. K. and Hwang J. (2014). *Introduction to probability,* Chapman & Hall/CRC.

Gentle, J. E., Hardle W., Mori Y. (2004). *Handbook of computational statistics*. Springer-Berlin.

Lange, K. (2010). *Numerical analysis for statisticians*, 2nd Edition, Springer, New York.

Pennoni, F. (2021). *Dispensa di Modelli Statistici II, parte di teoria e applicazioni con R*. Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca.

Rizzo M. L. (2008). *Statistical Computing with R*, Chapman & Hall/CRC, New York.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Semester

Semester I, cycle I, October-November 2021

Teaching language

The course is delivered in Italian. Erasmus students can use the didactic material in English and ask the teacher for the exam in English.